List of Data Projects

Python Automation Project & SQL Projects
By: Ain Najwa Arbain (Wawa)

1. Excel Data Extraction Automation

Python Automation Project
By: Ain Najwa Arbain (Wawa)
Year: 2022

Project Overview

- Excel workbooks are being widely used in the industry to store, capture and manipulate data.
- Manual work will need to be done in case of in need to copy data from 1 file to another.
- Manual data transfer can result in data inaccuracy, human error etc.
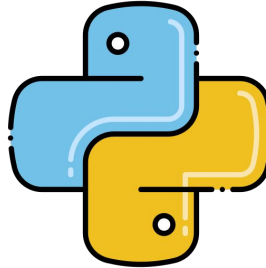
Solution

- Create a python automation project.
- Automate the process of data transfer from 1 workbook to another.
- Reduce the possibility of data inaccuracy, consistency and error.
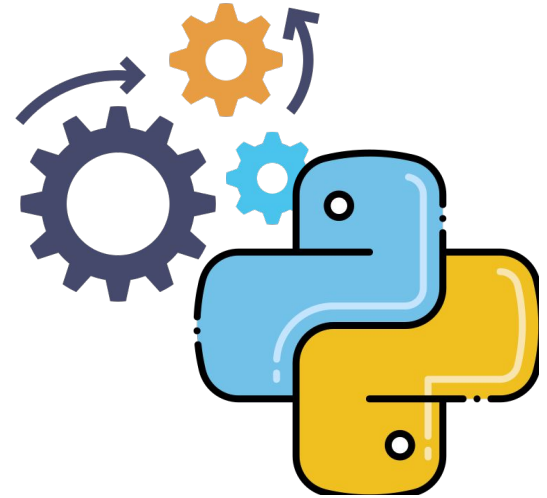
Automation Approach

Excel Spreadsheet

**Python**

New Excel Spreadsheet

- Open Folder A.
- Folders with a high number of Excel files.
- Each Excel file contains multiple worksheets

- The use of Python to automate the data extraction process.
- Iterating through each excel and each worksheet.
- Dynamically detect data bucket locations in excel.
- Store extracted data into dataframes.

- Automatically paste dataframes into a new excel file.
- Dynamically paste to its respective location.
- Save into Folder B.

Sample Data

- Create 2 folders in local machine, **Folder A** is considered to be the original folder with original excel files containing sample students data.

- While **Folder B** will be the designated folder of the new created excel files based on the extracted data.

Folder A

Folder B
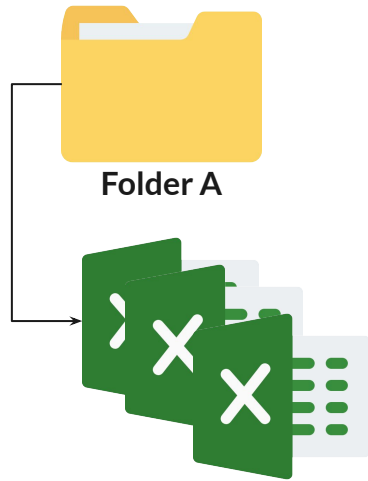
3 Excel Files containing 3 different classes and its students data
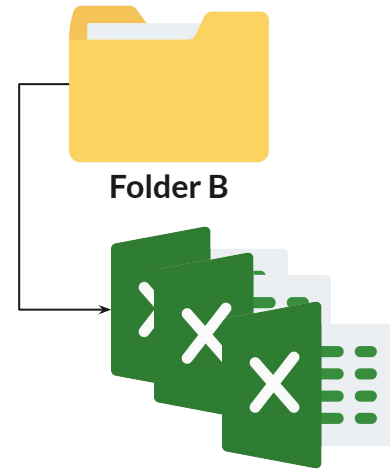
New Excel Files will be created containing summary data needed

Sample Data II

5 Mercury

Data Entry

First Name	Last Name	Class	Address	Contact	Emergency Contact Name	Relationship
Andrew	Garfield	5 Mercury	123, ABC Road Street, America	123456789	Ali	Uncle
Tom	Holland	5 Mercury	222, DDD Charlie Rd, America	222456098	Abu	Uncle
Anthony	Mackie	5 Mercury	223, DDD Charlie Rd Jr, America	123489208	Ah Cheng	Uncle
Robert	Downey Jr	5 Mercury	5434, Xannie Rd, America	209390897	Johnny	Father
Chris	Evans	5 Mercury	9090, OPQ Road Street, America	490274922	Chris Sr	Father
Chris	Hemsworth	5 Mercury	Level 23, First Road Condo, Australia	458304203	Emma	Mother
Sebastian	Stan	5 Mercury	Level 5, Tower Main Condo, America	450984039	Lilly	Mother

Data Entry

First Name	Last Name	Class	Emergency Contact Name	Relationship	Contact		Contact Option
Andrew	Garfield	5 Mercury	Ali	Uncle		123456789	Call
Tom	Holland	5 Mercury	Abu	Uncle		222456098	Call
Anthony	Mackie	5 Mercury	Ah Cheng	Uncle		123489208	Call
Robert	Downey Jr	5 Mercury	Johnny	Father		209390897	Call
Chris	Evans	5 Mercury	Chris Sr	Father		490274922	Call
Chris	Hemsworth	5 Mercury	Emma	Mother	NA		Email
Sebastian	Stan	5 Mercury	Lilly	Mother	NA		Email

Students Details | Students Performance By Terms | Students by Core Subject | ⊕

Dynamic Data Bucket Detection

```
Opening Workbook =  C:\Users\user\PycharmProjects\ExcelProject\Folder A\5 Mercury.xlsx

Current Active Sheet:  <Worksheet "Students Details">

Starting point data bucket =  [4, 14, 25]
Empty row in worksheet:  [13, 23, 24, 34]
Empty Column in worksheet:  [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]
```

- Dynamically detect initial and end data bucket location prior to creating dataframes
- In case user adds in more columns or rows, this code is expected to detect it properly

Dynamic Data Bucket Detection

```
range(4, 13)
            0            1   ...                            6             7
0  Data Entry          None  ...                         None          None
1        None    First Name  ...    Emergency Contact Name  Relationship
2        None        Andrew  ...                          Ali         Uncle
3        None           Tom  ...                          Abu         Uncle
4        None       Anthony  ...                     Ah Cheng         Uncle
5        None        Robert  ...                       Johnny        Father
6        None         Chris  ...                     Chris Sr        Father
7        None         Chris  ...                         Emma        Mother
8        None     Sebastian  ...                        Lilly        Mother

[9 rows x 8 columns]

range(14, 23)
            0            1            2   ...          5             6             7
0  Data Entry          None         None  ...       None          None          None
1        None    First Name    Last Name  ...  Relationship     Contact  Contact Option
2        None        Andrew     Garfield  ...       Uncle  123456789            Call
3        None           Tom      Holland  ...       Uncle  222456098            Call
4        None       Anthony       Mackie  ...       Uncle  123489208            Call
```

- Create dataframes based on the dynamic data bucket
- Range is detected by pairing technique.
- This will be iterated through all the worksheets within the workbook

Enhancement: Efficient Dataframe Creation

Data Entry

First Name	Last Name	Class	Emergency Contact Name	Relationship Contact	Contact Option
Andrew	Garfield	5 Mercury			
Tom	Holland	5 Mercury			
Anthony	Mackie	5 Mercury			
Robert	Downey Jr	5 Mercury			
Chris	Evans	5 Mercury			
Chris	Hemsworth	5 Mercury			
Sebastian	Stan	5 Mercury			

Why create dataframes if there is no data value? To eliminate this:

- % of null value is detected among each data buckets.
- If % exceeded a certain range, dataframes will not be created.
- This is to ensure much efficiency.

```
range(25, 34)


Process finished with exit code 0
```